On the Detection of Image-Scaling Attacks in Machine Learning

ACSAC'23, 08 Dec 2023

Erwin Quiring
International Computer Science Institute (ICSI), UC Berkeley
Ruhr-University Bochum

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Motivation

- Data preprocessing is often necessary for machine learning

Motivation

- Data preprocessing is often necessary for machine learning
- Computer Vision
  - Scaling necessary to match input dimensions
  - VGG19 expects $224 \times 224 \times 3$ pixels

Motivation

- Data preprocessing is often necessary for machine learning
- Computer Vision
  - Scaling necessary to match input dimensions
  - VGG19 expects $224 \times 224 \times 3$ pixels
- Natural Languages
  - Preprocess text to make it suitable for ML
  - Examples: Tokenization, lowercase conversion, stemming, stop word removal

Motivation

- Data preprocessing is often necessary for machine learning
- Computer Vision
  - Scaling necessary to match input dimensions
  - VGG19 expects $224 \times 224 \times 3$ pixels
- Natural Languages
  - Preprocess text to make it suitable for ML
  - Examples: Tokenization, lowercase conversion, stemming, stop word removal

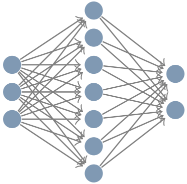
Unfortunately, preprocessing brings a new attack surface

$\longrightarrow$ "Adversarial Preprocessing" [Quiring et al., Usenix Sec'20]

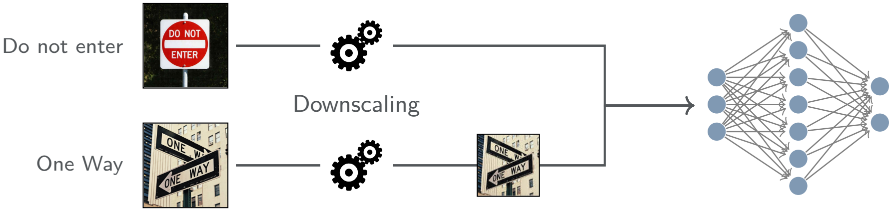
Motivation: Image Scaling

Do not enter

One Way



Motivation: Image Scaling



Motivation: Image Scaling

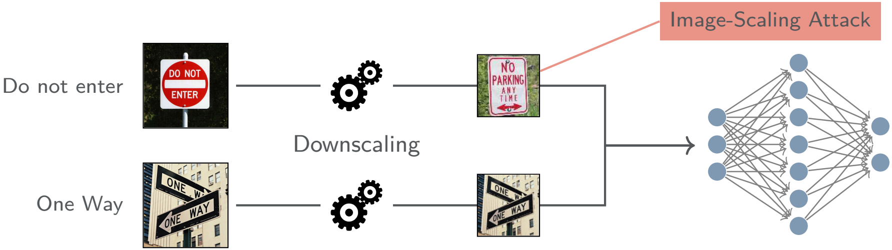


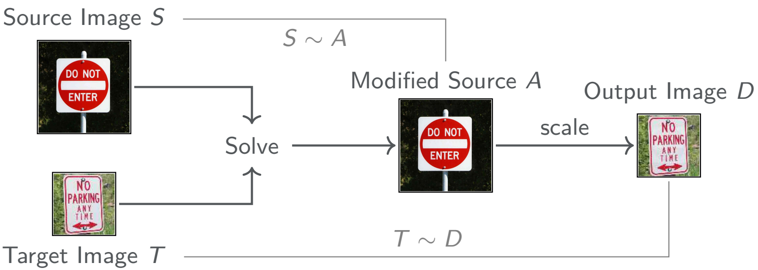
Image-Scaling Attacks

- Manipulated image changes appearance after downscaling

Xiao et al. 2019

Image-Scaling Attacks

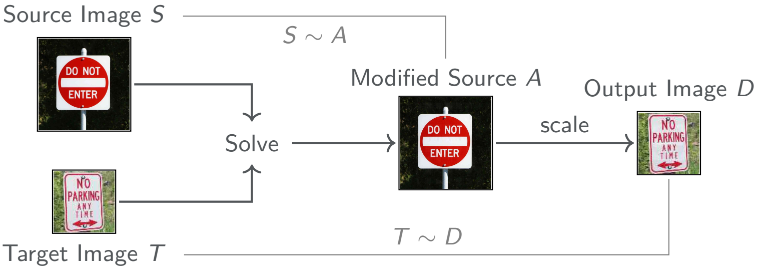
- Manipulated image changes appearance after downscaling



Xiao et al. 2019

Image-Scaling Attacks

- Manipulated image changes appearance after downscaling



- Both goals must be achieved: $T \simeq D$ and $S \simeq A$

Xiao et al. 2019

Threat Scenario in Machine Learning

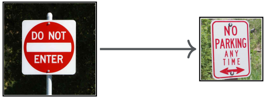
Possible attacks

Quiring and Rieck 2020, Xiao et al. 2019

Threat Scenario in Machine Learning

Possible attacks

- False predictions at test time

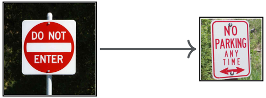


Quiring and Rieck 2020, Xiao et al. 2019

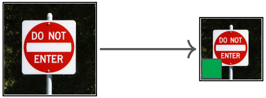
Threat Scenario in Machine Learning

Possible attacks

- False predictions at test time



- Conceal manipulations at training time

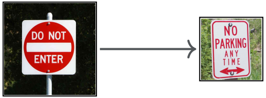


Quiring and Rieck 2020, Xiao et al. 2019

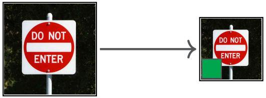
Threat Scenario in Machine Learning

Possible attacks

- False predictions at test time



- Conceal manipulations at training time



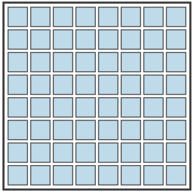
Capabilities and knowledge

- Attack is agnostic to learning model, data, features
- Knowledge of scaling algorithm only needed

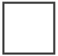
Quiring and Rieck 2020, Xiao et al. 2019

Root-Cause of Scaling Attack (Simplified)

Source Image S



Output Image D



Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

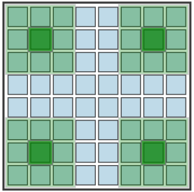
Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

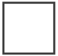
Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

Source Image S



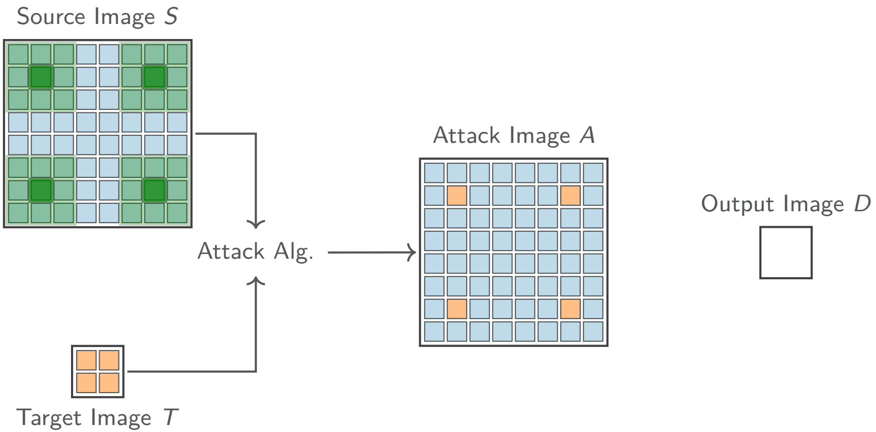
Output Image D



Target Image T

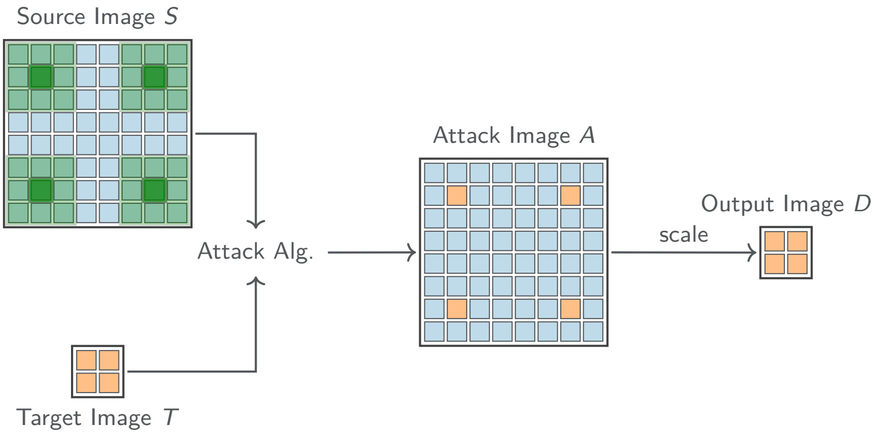
Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)



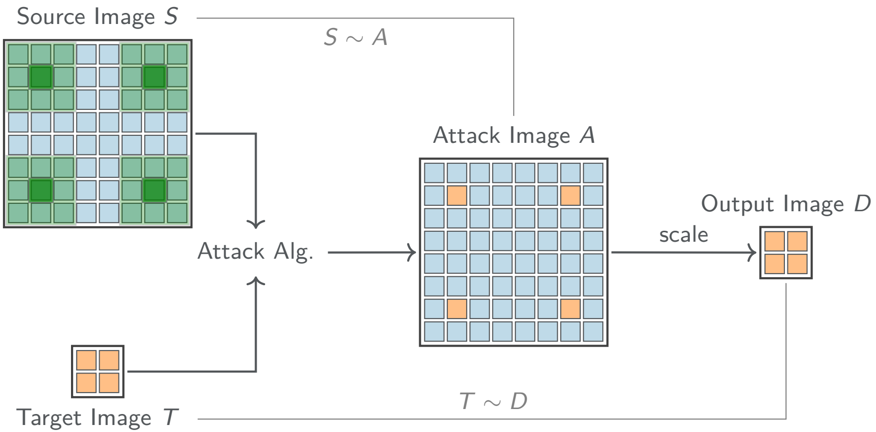
Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)



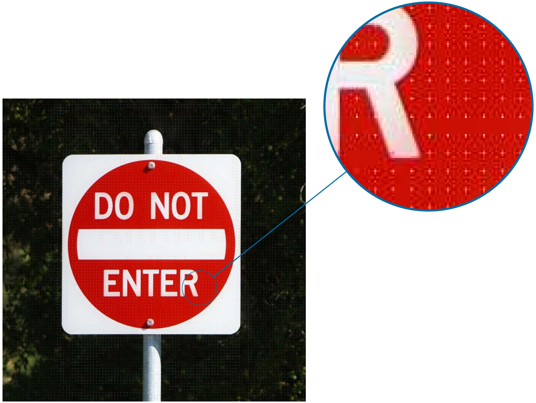
Quiring et al. 2020

Root-Cause of Scaling Attack (Simplified)

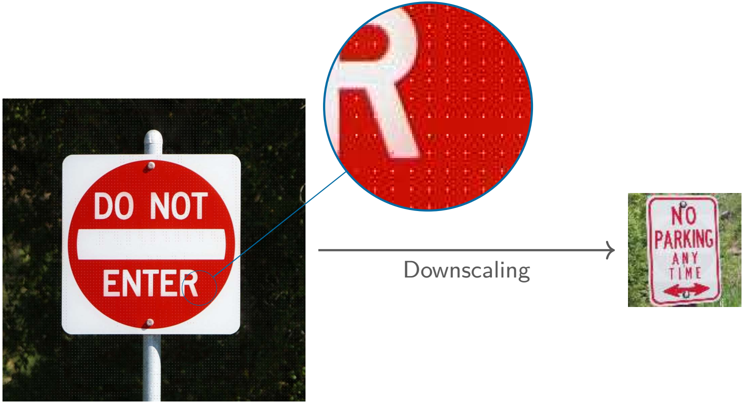


Quiring et al. 2020

Scaling Attack Example



Scaling Attack Example



Defenses

Two defense strategies

- **Prevention**: Implement robust scaling by design

Defenses

Two defense strategies

- **Prevention**: Implement robust scaling by design
- **Detection**: Find out that an attack is going on **[This work]**

  Reasons:
  - Be able to scan image collections or to identify attacker
  - Robust scaling is slower than vulnerable scaling
  - Detection necessary for proprietary learning systems

Paper Contribution: Detection of Scaling Attacks

- First in-depth systematization and analysis of detection

Paper Contribution: Detection of Scaling Attacks

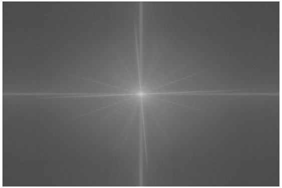
- First in-depth systematization and analysis of detection
- Two general paradigms identified
- Novel detection methods for these paradigms

General Paradigm 1: Frequency Analysis

- **Paradigm**: Analyze frequency spectrum of image
- Frequency representation shows periodical patterns

General Paradigm 1: Frequency Analysis

- **Paradigm**: Analyze frequency spectrum of image
- Frequency representation shows periodical patterns

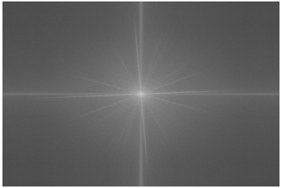


(a) Unmodified image S

General Paradigm 1: Frequency Analysis

- **Paradigm**: Analyze frequency spectrum of image
- Frequency representation shows periodical patterns

- Recall root cause: Attack injects pixels in periodic distance
- Thus, attack causes unique, periodic frequency peaks

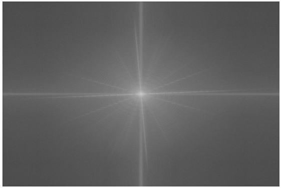


(a) Unmodified image S

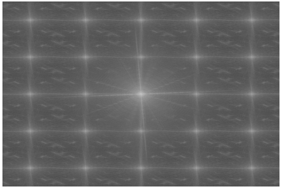
General Paradigm 1: Frequency Analysis

- **Paradigm**: Analyze frequency spectrum of image
- Frequency representation shows periodical patterns

- Recall root cause: Attack injects pixels in periodic distance
- Thus, attack causes unique, periodic frequency peaks



(a) Unmodified image S



(b) Attack image A

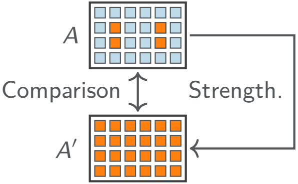
General Paradigm 2: Spatial Analysis

- **Paradigm**: Analyze pixels of image
- Naturally advantage of knowing (potentially modified) scaling pixels

General Paradigm 2: Spatial Analysis

- **Paradigm**: Analyze pixels of image
- Naturally advantage of knowing (potentially modified) scaling pixels

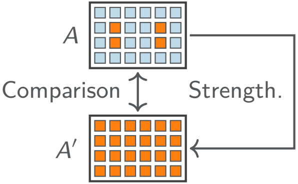
Variant 1: Adversarial-Signal Driven



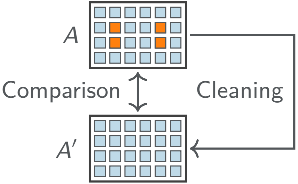
General Paradigm 2: Spatial Analysis

- **Paradigm**: Analyze pixels of image
- Naturally advantage of knowing (potentially modified) scaling pixels

Variant 1: Adversarial-Signal Driven



Variant 2: Clean-Signal Driven



Paper Contribution: Detection of Scaling Attacks (Continued)

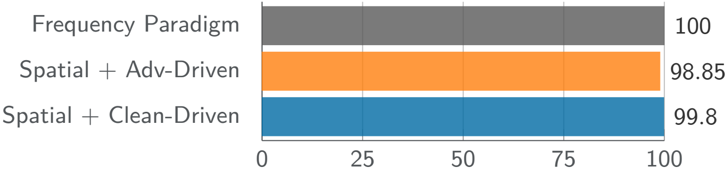
- First in-depth systematization and analysis of detection
- Two general paradigms identified
- Novel detection methods for these paradigms

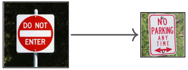
Paper Contribution: Detection of Scaling Attacks (Continued)

- First in-depth systematization and analysis of detection
- Two general paradigms identified
- Novel detection methods for these paradigms
- Comprehensive evaluation
  - In total, 19 detection methods compared
  - Diverse modification scenarios (full, overlay, or just local image modification)
  - Varying setups:
    - Multiple learning platforms & scaling algorithms
    - ImageNet with varying images & scaling ratios
    - Static & adaptive adversaries

Some Key Results (1/2)

Global scenario

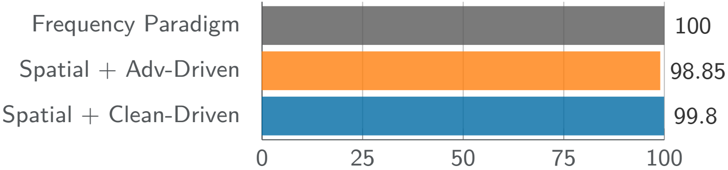


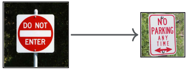


Local scenario

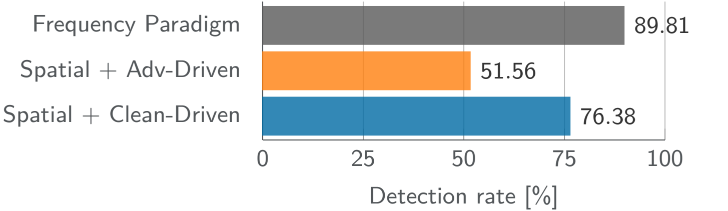
Some Key Results (1/2)

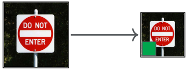
Global scenario





Local scenario





Some Key Results (2/2)

Under static attackers:

- Frequency paradigm is excellent in global & local scenario
- Spatial paradigm is excellent in global, and satisfactory in local scenario

Some Key Results (2/2)

Under static attackers:

- Frequency paradigm is excellent in global & local scenario
- Spatial paradigm is excellent in global, and satisfactory in local scenario

Under adaptive attackers:

- Frequency paradigm is vulnerable
- Spatial paradigm provides robustness

Some Key Results (2/2)

Under static attackers:

- Frequency paradigm is excellent in global & local scenario
- Spatial paradigm is excellent in global, and satisfactory in local scenario

Under adaptive attackers:

- Frequency paradigm is vulnerable
- Spatial paradigm provides robustness

Recommendation: Use Ensemble

Summary

- Data preprocessing is essential in ML
- Unfortunately, it leads to a new attack surface

Summary

- Data preprocessing is essential in ML
- Unfortunately, it leads to a new attack surface

- *"Adversarial Preprocessing"* relevant in multiple domains
  - Image-scaling attacks in computer vision
  - NLP attacks

    See *No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning*, Eisenhofer, Quiring, et al., Usenix Sec'23

Summary

- Data preprocessing is essential in ML
- Unfortunately, it leads to a new attack surface
- *"Adversarial Preprocessing"* relevant in multiple domains
  - Image-scaling attacks in computer vision
  - NLP attacks

    See *No more Reviewer #2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning*, Eisenhofer, Quiring, et al., Usenix Sec'23
- This work:
  - First in-depth systematization and analysis of detection
  - Efficient detection of scaling attacks possible
  - Ensemble of varying paradigms necessary

References I

[1] Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. "Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning". In: *Proc. of USENIX Security Symposium*. 2020.

[2] Erwin Quiring and Konrad Rieck. "Backdooring and Poisoning Neural Networks with Image-Scaling Attacks". In: *Deep Learning and Security Workshop (DLS)*. 2020.

[3] Qixue Xiao, Yufei Chen, Chao Shen, Yu Chen, and Kang Li. "Seeing is Not Believing: Camouflage Attacks on Image Scaling Algorithms". In: *Proc. of USENIX Security Symposium*. 2019.